Predicting the fixation density over time

Heiko H. Schütt, Lars O. M. Rothkegel, Hans A. Trukenbrod,
Ralf Engbert & Felix A. Wichmann

When modelling human eye movements we usually separate bottom-up and top-down effects, i.e. whether some effect is caused by the stimulus or by some internal state of the observer like their tasks or intentions. We also separate the orthogonal dimension whether the features used to guide eye movements are low-level—like local contrast, luminance or orientation content—or high-level—like object locations, scene congruence or scene category. Furthermore, humans display systematic tendencies in their eye movements, preferring certain saccade lengths and directions and a tendency to continue moving into the same direction.

It is unclear how these different factors interact to determine were we look and most models only include a small selection of the mentioned influence factors. To disentangle them, we analyse the dependencies between fixations and the fixation densities over time. In our analysis we include how well fixation densities are predicted by, first, low-level bottom-up saliency, including a saliency model based on our early spatial vision model (Schütt and Wichmann, 2017), and, second, a recent DNN-based saliency model including low- and high-level bottom-up saliency (DeepGaze II, Kümmerer et al., 2016).

To separate top-down effects, we use two datasets: One Corpus dataset in which 105 subjects looked at 90 images to memorize them and a search dataset in which 10 subjects searched for 6 different targets with varying spatial frequency and orientation content superimposed over 25 natural images 8 times each resulting in 480 searches per image.

Based on the Corpus Dataset we separate the exploration into three phases: An onset response with the first saccade, an initial exploration lasting around 10 fixations and a final equilibrium phase. First fixations are most predictable but follow a different density than later ones. During the initial exploration fixations gradually become less predictable. Finally, the fixation density stops broadening and the equilibrium state is reached in which fixations are least focussed but still favour the same areas as during the exploration.

The prediction quality of all saliency models follows the curve of predictable information. They predict fixations best at the beginning and gradually get worse. The simple saliency model based on our early spatial vision model performs as well as classical saliency models. However, DeepGaze II performs substantially better by using high-level information throughout the whole trial. This advantage is present at the latest 200 ms after image onset; however, the predictive power of the early vision saliency model for the first fixation(s) is much better than for later fixations, and as a corollary the advantage of DeepGaze II is relatively small for the first fixation(s).

On the search dataset all saliency models perform badly after a small initial prediction success, even if the non-linearity and central fixation bias are newly adjusted to the search data. Instead we observe that subjects adjust where they look and their eye movement dynamics to the target they search for. Specifically they make shorter saccades and exhibit shorter fixation times for higher frequency targets than for lower frequency targets.

Our observations confirm that bottom-up guidance of eye movements can be overwritten almost entirely by task effects in static natural scenes. Nonetheless our data support some early bottom-up guidance, which includes high-level features already for the very first saccade.

References

Kümmerer, M., Wallis, T. S. A., and Bethge, M. (2016). DeepGaze II: Reading fixations from deep features trained on object recognition. *arXiv:1610.01563 [cs, q-bio, stat]*.

Schütt, H. H. and Wichmann, F. A. (2017). An image-computable psychophysical spatial vision model. *Journal of Vision*, 17(12):12:1–35.

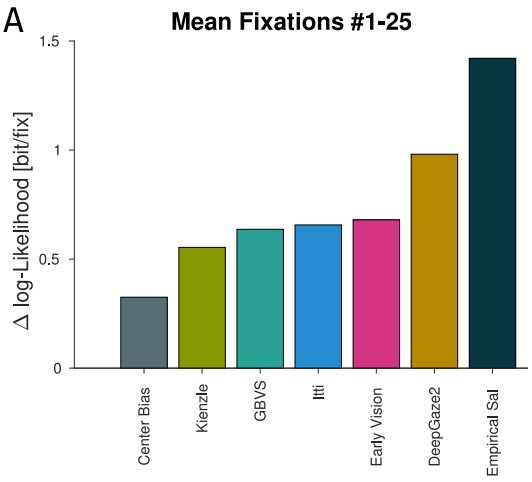

Figure 1: Overall saliency model performance on the Corpus dataset. A simple weighted sum of a early visual representation performs slightly better than classical saliency models, but DeepGaze II performs substantially better using high-level information.

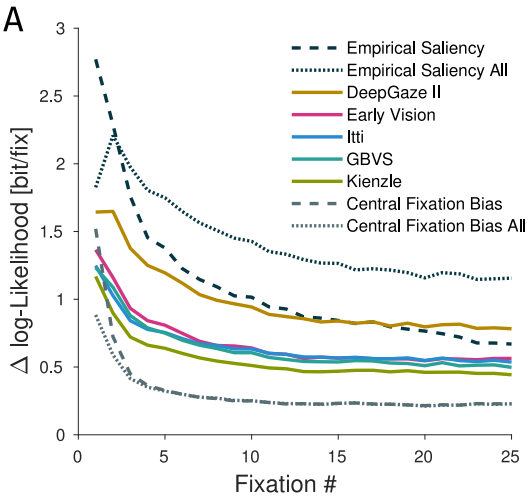

Figure 2: Performance of the saliency models over time. Empirical saliency and Central fixation bias were estimated as kernel density estimates with leave one subject out crossvalidation based on either fixations with the same number or all fixations except for the first.

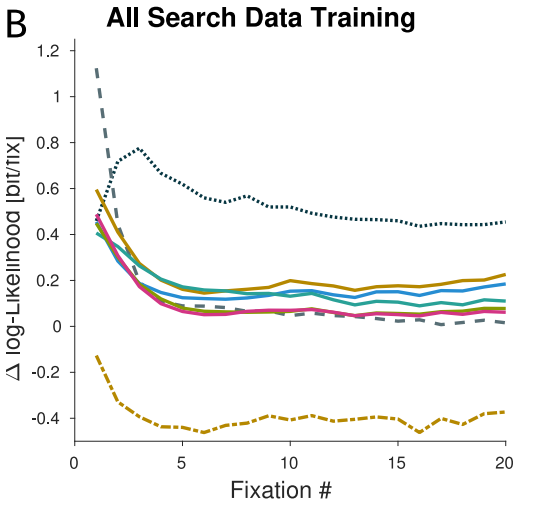

Figure 3: Saliency model performance on the Search dataset. The dot-dashed line represents the performance of the unadjusted DeepGaze II model prediction.